# Testing for equality between two transformations of random variables

Mohamed Boutahar, Denys Pommeret

# *Testing for equality between two transformations of random variables*

Mohamed BOUTAHAR **and Denys POMMERET † †

October 31, 2011

**Abstract**

Consider two random variables contaminated by two unknown transformations. The aim of this paper is to test the equality of those transformations. Two cases are distinguished: first, the two random variables have known distributions. Second, they are unknown but observed before contaminations. We propose a nonparametric test statistic based on empirical cumulative distribution functions. Monte Carlo studies are performed to analyze the level and the power of the test. An illustration is presented through a real data set.

*Keywords* : empirical cumulative distribution; nonlinear contamination; nonparametric estimation

## 1 Introduction

There exists an important literature concerning the deconvolution problem, when an unknown signal $Y$ is contaminated by a noise $Z$, leading to the observed signal

$$X = Y + Z. \tag{1}$$

A major problem is to reconstruct the density of $Y$. Many authors studied the univariate problem when the noise $Z$ has known distribution (see for instance Fan [10], Carroll and Hall [3], Devroye [7], or more recently Holzmann *et al.* [12] for a review). Bissantz *et al.* [1] proposed the construction of confidence bands for the density of $Y$ based on i.i.d. observations from (1). The case where both $Y$ and $Z$ have unknown distributions is considered in Neumann [15], Diggle and Hall [8] or Johannes *et al.* [13] among others. When the error density and the distribution of $Y$ have different characteristics the model can be identified as shown in Butucea and Matias [2] and Meister [14]. But without information

**Corresponding author, IML. Luminy Faculty of Sciences. 163 Av. de Luminy 13288 Marseille Cedex 9 e-mail: boutahar@univmed.fr.

†† IML. Luminy Faculty of Sciences. 163 Av. de Luminy 13288 Marseille Cedex 9 e-mail: pommeret@univmed.fr.

on $Z$, the model suffers of identification conditions. One solution is to assume another independent sample is observed from the measurement error $Z$ (as done in Efromovich and Koltchinskii [9] and Cavalier and Hengartner [5]).

A more general model than (1) occurs when the contaminated random variables are observed through a transformation; that is, there exists $g$ such that

$$X = g(Y + Z). \tag{2}$$

When $g$ is known the problem is to estimate the distribution of $Y$, observing a sample from (2). An application of this model to fluorescence lifetime measurements is given in Comte and Rebafka [6]. The authors developed an adaptative estimator that take into account the perturbation from the unknown additive noise, and the distortion due to the nonlinear transformation.

In this paper we consider a two sample problem of contamination that can be related to models (1) and (2) as follows: We assume that two contaminated random variables are observed, say $X$ and $\tilde{X}$, which are transformations of two known, or observed, signals, that is:

$$X = g(Y), \qquad \tilde{X} = \tilde{g}(\tilde{Y}), \tag{3}$$

where $g$ and $\tilde{g}$ are continuous monotone unknown functions. Our purpose is to test

$$H_0 : g = \tilde{g} \quad \text{against} \quad H_1 : g \neq \tilde{g}, \tag{4}$$

based on two i.i.d. samples satisfying (3). The problem of testing (4) is of interest in many applications when a signal is noised in another way than the additive noise model (1). We will distinguish two important cases:

*Case 1* The distributions of $Y$ and $\tilde{Y}$ are known and we observe two samples reflecting $X$ and $\tilde{X}$. This situation may be encountered when two signals are controlled in entry but observed with perturbations in exit of a system.

*Case 2* The distributions of $Y$ and $\tilde{Y}$ are unknown and we first observe two independent samples based on $Y$ and $\tilde{Y}$, and then we observe contaminated samples $X$ and $\tilde{X}$ satisfying (3). This situation may be encountered when two unknown signals are observed both in entry and in exit of a system.

For both cases we construct a test statistics based on non parametric empirical estimators of $g$ and $\tilde{g}$ and we adapt a limit result on empirical processes due to Sen [16]. Our test statistics are very easily implemented and we observe through simulations that they have a good power against various alternatives. It is clear that when $H_0$ is not rejected; that is when the two noise functions are identical, it is then of interest to interpret the common estimation of $g$. We illustrate this point with a study of the Framingham dataset (see Carroll *et al.* [4], and more recently Wang and Wang [17]).

The paper is organized as follows: in Section 1 we consider the problem when the two original signals have known distributions. In Section 2 we relax the last assumption by assuming unknown distributions but we observe the two original signals after and before perturbations. In Section 3 a simulation study is presented and a real data set is analyzed.

## 2 The test statistic

### 2.1 Case 1: the two signal distributions of $Y$ and $\tilde{Y}$ are known

We consider $n$ (resp. $\tilde{n}$) i.i.d. observations $X_1, \cdots, X_n$ (resp. $\tilde{X}_1, \cdots, \tilde{X}_{\tilde{n}}$) from (3). We assume that $Y$ and $\tilde{Y}$ are independent. Write $F_Y$ and $F_{\tilde{Y}}$ the cumulative distribution functions of $Y$ and $\tilde{Y}$ respectively. We assume that these functions are known and invertible. We also write $F_X$ and $F_{\tilde{X}}$ the cumulative distribution functions of $X$ and $\tilde{X}$. Also we assume that the transformations $g$ and $\tilde{g}$ are monotone and, without loss of generality, that they are increasing. Note that $g(y) = F_X^{-1}(F_Y(y))$ and $\tilde{g}(y) = F_{\tilde{X}}^{-1}(F_{\tilde{Y}}(y))$. Hence a natural nonparametric estimators of the contaminating functions are given by

$$\widehat{g}(\cdot) = X_{([nF_Y(\cdot)]+1)} \quad \text{and} \quad \widehat{\tilde{g}}(\cdot) = \tilde{X}_{([\tilde{n}F_{\tilde{Y}}(\cdot)]+1)}, \tag{5}$$

where $X_{(i)}$ and $\tilde{X}_{(i)}$ denote the $i$th order statistics, and $[x]$ denotes the integer part of the real $x$. A fundamental theorem of Sen [16] states the following convergence in distribution

$$\sqrt{n}\big(X_{([np]+1)} - F_X^{-1}(p)\big) \xrightarrow{D} \mathcal{N}\left(0, \frac{p(1-p)}{f^2(F_X^{-1}(p))}\right), \ \forall p \in (0,1), \tag{6}$$

where $\xrightarrow{D}$ denotes the convergence in distribution, $f$ denotes the density of $X$ and $\mathcal{N}(m, \sigma^2)$ the Normal distribution with mean $m$ and variance $\sigma^2$. We will need the following two standard assumptions:

- $(A_1)$ there exists $a < \infty$ such that $n/(n+\tilde{n}) \to a$
- $(A_2)$ $f > 0$ and $f$ is $\mathcal{C}^k$, for some positive integer $k$.

We deduce a first result which is a main tool for the construction of the test statistic.

PROPOSITION 2.1 Let Assumption $(A_1) - (A_2)$ hold. Under $H_0$ we have

$$\sqrt{\frac{n\tilde{n}}{n+\tilde{n}}}\left(\widehat{g}(y) - \widehat{\tilde{g}}(y)\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2(y)), \text{ as } n \to \infty, \tilde{n} \to \infty, \tag{7}$$

where

$$\sigma^2(y) = (1-a)\frac{F_Y(y)(1-F_Y(y))}{f_X^2(g(y))} + a\frac{F_{\tilde{Y}}(y)(1-F_{\tilde{Y}}(y))}{f_{\tilde{X}}^2(\tilde{g}(y))}$$

PROOF. It follows directly from (6), replacing $p$ by $F_Y(y)$ and $F_{\tilde{Y}}(y)$ respectively. ■

We will estimate the variance $\sigma^2$ by using a nonparametric method. Consider a kernel $K(\cdot)$, for instance the quartic kernel defined by $K(y) = \frac{15}{16}(1-y^2)^2\mathbf{1}_{(-1,1)}(y)$, and an associated bandwidth $h_n$. In the sequel, we will set $K_{h_n}(y) = K(\frac{y}{h_n})$. To avoid small values for denominators in the estimation of the variance we use

$$\widehat{f}_X(y) \quad = \quad \max\left(\frac{1}{nh_n}\sum_{i=1}^{n} K_{h_n}(X_i - y), e_n\right)$$

and

$$\widehat{f}_{\tilde{X}}(y) = \max\left(\frac{1}{\tilde{n}h_{\tilde{n}}}\sum_{i=1}^{\tilde{n}} K_{h_{\tilde{n}}}(\tilde{X}_i - y), e_{\tilde{n}}\right),$$

where $e_n > 0$ and $e_n \to 0$ when $n$ tends to infinity. The estimator of $\sigma^2$ is then

$$\widehat{\sigma}^2(y) \quad = \quad (1-a)\frac{F_Y(y)(1-F_Y(y))}{\widehat{f}^2_X(\widehat{g}(y))} + a\frac{F_{\tilde{Y}}(y)(1-F_{\tilde{Y}}(y))}{\widehat{f}^2_{\tilde{X}}(\widehat{\tilde{g}}(y))},$$

and we consider the statistic

$$T_1(y) \quad = \quad \frac{n\tilde{n}}{n+\tilde{n}}\widehat{\sigma}(y)^{-2}\left(\widehat{g}(y) - \widehat{\tilde{g}}(y)\right)^2. \tag{8}$$

PROPOSITION 2.2 Let Assumptions $(A_1)$-$(A_2)$ hold. If $h_n \simeq n^{-c_1}$, $e_n \simeq n^{-c_2}$ for some positive constants $c_1$ and $c_2$ such that $\frac{c_2}{k} < c_1 < \frac{1}{1+2k}$, then under $H_0$, when $n \to \infty, \tilde{n} \to \infty$, we have for all $y$:

$$T_1(y) \xrightarrow{D} Z,$$

where $Z$ is Chi-squared distributed with one degree of freedom.

PROOF. We need the fundamental Lemma (see Härdle [11]):

LEMMA 2.1

$$\sup_{y\in\mathbb{R}} |\hat{f}^2(y) - f^2(y)| = \mathcal{O}_p\left(h_n^{2k} + \frac{\log n}{nh_n}\right).$$

We can write

$$\widehat{\sigma}^2(y) \quad = \quad \frac{u(y)}{\widehat{f}^2_X(\widehat{g}(y))} + \frac{v(y)}{\widehat{f}^2_{\tilde{X}}(\widehat{\tilde{g}}(y))},$$

where $u(y) = (1-a)F_Y(y)(1-F_Y(y))$ and $v(y) = aF_{\tilde{Y}}(y)(1-F_{\tilde{Y}}(y))$. Using Taylor expansion there exist $A$ and $B$ such that

$$\widehat{\sigma}^2(y) \quad = \quad \sigma^2(y) + \left(\widehat{f}^2_X(\widehat{g}(y)) - f^2(g(y))\right)\left(\frac{-1}{A^2}\right) + \left(\widehat{\tilde{f}}^2_X(\widehat{\tilde{g}}(y)) - \tilde{f}^2(\tilde{g}(y))\right)\left(\frac{-1}{B^2}\right),$$

with

$$\frac{1}{A^2} \leq \frac{1}{e_n^2} \qquad \text{and} \quad \frac{1}{B^2} \leq \frac{1}{e_{\tilde{n}}^2}.$$

Then, from Lemma 2.1 we get

$$\widehat{\sigma}^2(y) \quad = \quad \sigma^2(y) + o_P(1),$$

by assumption and the result follows from Proposition 2.1.

■

## 2.2 Case 2: the two signal distributions $Y$ and $\tilde{Y}$ are unknown

We consider $n_x$ (resp. $\tilde{n_x}$) i.i.d. observations $X_1, \cdots, X_{n_x}$ (resp. $\tilde{X}_1, \cdots, \tilde{X}_{\tilde{n_x}}$) and $n_y$ (resp. $\tilde{n_y}$) i.i.d. observations $Y_1, \cdots, Y_{n_y}$ (resp. $\tilde{Y}_1, \cdots, \tilde{Y}_{\tilde{n_y}}$) from (3). Put

$$N = n_x n_y/(n_x + n_y) \quad \text{and} \quad \tilde{N} = \tilde{n_x}\tilde{n_y}/(\tilde{n_x} + \tilde{n_y}).$$

The two samples $Y_1, \cdots, Y_{n_y}$ and $\tilde{Y}_1, \cdots, \tilde{Y}_{\tilde{n_y}}$ can be viewed as two independent training sets which permit to estimate the initial densities of the signals before perturbations. Again we want test $H_0 : g = \tilde{g}$. We now estimate $g$ and $\tilde{g}$ by

$$\widehat{g}(\cdot) = X_{([n_x \widehat{F}_Y(\cdot)])} \quad \text{and} \quad \widehat{\tilde{g}}(\cdot) = \tilde{X}_{([\tilde{n}_x \widehat{F}_{\tilde{Y}}(\cdot)])}, \tag{9}$$

where

$$\widehat{F}_Y(y) = \frac{1}{n_y}\sum_{i=1}^{n_y} \mathbf{1}_{\{Y_i \le y\}} \quad \text{and} \quad \widehat{F}_{\tilde{Y}}(y) = \frac{1}{\tilde{n}_y}\sum_{i=1}^{\tilde{n}_y} \mathbf{1}_{\{\tilde{Y}_i \le y\}}, \tag{10}$$

are the empirical distribution functions of $Y$ and $\tilde{Y}$ respectively. We assume that

$$\lim n_x/(n_x + n_y) = a < \infty, \qquad \lim \tilde{n_x}/(\tilde{n_x} + \tilde{n_y}) = \tilde{a} < \infty,$$

and we make the following assumption, extending Assumption (A1):

- $(A_3)$ there exists $b < \infty$ such that $N/(N + \tilde{N}) \to b$.

We can extend Proposition 2.1 as follows:

PROPOSITION 2.3 Let Assumption $(A_2) - (A_3)$ hold. Under $H_0$ we have

$$\sqrt{\frac{N\tilde{N}}{N + \tilde{N}}}\left(\widehat{g}(y) - \widehat{\tilde{g}}(y)\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2(y)), \text{ as } N \to \infty, \tilde{N} \to \infty, \tag{11}$$

where

$$\sigma^2(y) = (1-b)\frac{F_Y(y)(1 - F_Y(y))}{f_X^2(g(y))} + b\frac{F_{\tilde{Y}}(y)(1 - F_{\tilde{Y}}(y))}{f_{\tilde{X}}^2(\tilde{g}(y))}. \tag{12}$$

PROOF. We first show that

$$U = \sqrt{\frac{n_x n_y}{n_x + n_y}}\left(\widehat{g}(y) - g(y)\right) \xrightarrow{D} \mathcal{N}(0, \sigma_1^2(y)), \text{ as } n_x \to \infty, n_y \to \infty,$$

where

$$\sigma_1^2(y) = \frac{F_Y(y)(1 - F_Y(y))}{f_X^2(g(y))}.$$

For that write

$$\widehat{g}(y) - g(y) \quad = \quad \widehat{G}(y) + G(y),$$

where $\widehat{G}(y) = \widehat{g}(y) - F_X^{-1}(\widehat{F}_Y(y)) = X_{([n_x \widehat{F}_Y(x)])} - F_X^{-1}(\widehat{F}_Y(y))$ and $G(y) = F_X^{-1}(\widehat{F}_Y(y)) - g(y)$. By the delta method we get

$$n_y^{1/2} G(y) \quad \to \quad \mathcal{N}(0, \sigma_1^2(y)).$$

Then we decompose the characteristic function

$$E\left(e^{iuU}\right) \quad = \quad E\left(e^{iun_{x,y}G} E(e^{iun_{x,y}\widehat{G}} | \mathbf{Y})\right),$$

where $n_{x,y} = \sqrt{\dfrac{n_x n_y}{n_x + n_y}}$ and $\mathbf{Y}$ stands for the vector of observation $Y_1, \cdots, Y_{n_y}$.
Since these functions are bounded we get:

$$\begin{aligned}
\lim_{n_x \to \infty, n_y \to \infty} E(\exp(iuU)) \quad &= \quad E\left(\lim_{n_x \to \infty, n_y \to \infty} e^{iun_{x,y}G} \lim_{n_x \to \infty, n_y \to \infty} E\left(e^{iun_{x,y}\widehat{G}} | \mathbf{Y}\right)\right) \\
&= \quad E\left(e^{iu\sqrt{a}Z} \lim_{n_y \to \infty} e^{-\frac{1}{2}(1-a)\widehat{\sigma}_1^2(y)}\right),
\end{aligned}$$

where $Z \sim \mathcal{N}(0, \sigma_1^2(y))$ and $\widehat{\sigma}_1^2(y) = \frac{\widehat{F}_Y(y)(1-\widehat{F}_Y(y))}{f_X^2(g(y))}$. We finally obtain

$$\lim_{n_x \to \infty, n_y \to \infty} E(\exp(iuU)) \quad = \quad \exp(-1/2 u^2 \sigma_1^2(y)).$$

Similarly, writing

$$\tilde{U} \quad = \quad \sqrt{\frac{\tilde{n}_x \tilde{n}_y}{\tilde{n}_x + \tilde{n}_y}} \left(\widehat{\tilde{g}}(y) - \tilde{g}(y)\right),$$

we obtain that

$$\tilde{U} \quad \stackrel{D}{\to} \quad \mathcal{N}(0, \tilde{\sigma}_1^2(y)), \text{ as } \tilde{n}_x \to \infty, \tilde{n}_y \to \infty,$$

with

$$\tilde{\sigma}_1^2(y) = \frac{F_{\tilde{Y}}(y)(1 - F_{\tilde{Y}}(y))}{f_{\tilde{X}}^2(\tilde{g}(y))}.$$

Finally, combining these two convergences with the equality $\tilde{g} = g$ under $H_0$ we complete the proof.

■

As previously we can estimate $\sigma^2(y)$ in (12) by a nonparametric estimator

$$\widehat{\sigma}^2(y) \quad = \quad (1-b)\frac{\widehat{F}_Y(y)(1 - \widehat{F}_Y(y))}{\widehat{f}_X^2(\widehat{g}(y))} + b\frac{\widehat{F}_{\tilde{Y}}(y)(1 - \widehat{F}_{\tilde{Y}}(y))}{\widehat{f}_{\tilde{X}}^2(\widehat{\tilde{g}}(y))},$$

where $\widehat{F}_Y$ and $\widehat{F}_{\tilde{Y}}$ are the empirical distribution functions of $Y$ and $\tilde{Y}$ given by (10). Our test statistic is given by

$$T_2(y) \quad = \quad \frac{N\tilde{N}}{N+\tilde{N}}\widehat{\sigma}^{-2}(y)\left(\widehat{g}(y)-\widehat{\tilde{g}}(y)\right)^2. \tag{13}$$

We can now generalize Proposition 2.2 as follows.

PROPOSITION 2.4 Let Assumptions $(A_2)$-$(A_3)$ hold. If $h_n \simeq n^{-c_1}$, $e_n \simeq n^{-c_2}$ for some positive constants $c_1$ and $c_2$ such that $\frac{c_2}{k} < c_1 < \frac{1}{1+2k}$, then under $H_0$, when $N \to \infty, \tilde{N} \to \infty$, we have:

$$T_2 \xrightarrow{D} Z,$$

where $Z$ is Chi-squared distributed with one degree of freedom.

PROOF. We combine the proof of Proposition 2.1 with the fact that $\widehat{F}(1-\widehat{F})$ is bounded to get

$$\widehat{\sigma}^2(y) \quad = \quad \sigma^2(y) + o_P(1),$$

and we conclude by Proposition 2.3. ■

## 2.3 Behaviour of the tests under $H_1$

We study convergence properties of the tests $T_1$ and $T_2$ under some alternatives

PROPOSITION 2.5

*a. General alternatives.*
*Consider the test statistics $T_1$ and $T_2$, then for all $y$ such that $g(y) \neq \tilde{g}(y)$, we have*

$$T_1(y) \quad \xrightarrow{P} +\infty \text{ and } T_2(y) \xrightarrow{P} +\infty,$$

*where $\xrightarrow{P}$ denotes the convergence in probability.*

*b. Local alternatives.*
*Let us denote $m = \frac{n\tilde{n}}{n+\tilde{n}}$ or $m = \frac{N\tilde{N}}{N+\tilde{N}}$ according to whether if the test statistic $T_1$ or $T_2$ is used and consider the local alternatives*

$$H_{l1} : \tilde{g}(y) = g(y) + \frac{k(y)}{m^\beta},$$

*then under $H_{l1}$ and when $n \to \infty$, $\tilde{n} \to \infty$, $N \to \infty$, $\tilde{N} \to \infty$ we have for all $y$:*
*i. If $\beta > 1/2$ then*

$$T_1(y) \xrightarrow{D} Z \text{ and } T_2(y) \xrightarrow{D} Z$$

*ii. If* $\beta = 1/2$ *then*

$$T_1(y) \xrightarrow{D} Z_k \text{ and } T_2(y) \xrightarrow{D} Z_k$$

*iii. If* $\beta < 1/2$ *then*

$$T_1(y) \xrightarrow{P} +\infty \text{ and } T_2(y) \xrightarrow{P} +\infty$$

*where* $Z$ *is Chi-squared distributed with one degree of freedom and* $Z_k$ *is a decentred Chi-squared distributed with one degree of freedom and parameter* $k(y)$.

The proof of this proposition is straightforward and hence is omitted.

REMARK 2.1 *Estimators* $\widehat{g}$ *(resp.* $\widehat{\tilde{g}}$ *) are computed from* $(X_1, \cdots, X_{n_x})$ *and* $(Y_1, \cdots, Y_{n_y})$ *(resp.* $(\tilde{X}_1, \cdots, \tilde{X}_{\tilde{n_x}})$ *and* $(\tilde{Y}_1, \cdots, \tilde{Y}_{\tilde{n_y}})$ *). Under the null* $H_0$ *there are two different ways to construct a common estimator of* $g$. *First we can consider the aggregate estimator*

$$\widehat{g}_0 = \frac{(n_x + n_y)\widehat{g} + (\tilde{n_x} + \tilde{n_y})\widehat{\tilde{g}}}{n_x + n_y + \tilde{n_x} + \tilde{n_y}}, \tag{14}$$

*and, second, another estimator can be construct by aggregating the samples.*

# 3 Simulations and data study

For all empirical powers or empirical levels we carry out experiments of 10000 samples and we use three different sample sizes: $n = 50$, $n = 100$, and $n = 500$. For each replication we compute the statistics $T_1(y)$ and $T_2(y)$ given by (8) and (13), where $y$ is chosen randomly following a standard normal distribution.

## 3.1 Study of the empirical levels

We will denote by $\mathcal{N}(0, 1)$ the standard normal distribution with mean zero and variance 1. We first consider the case where $Y_t$ and $\tilde{Y}_t$ are independent and $\mathcal{N}(0, 1)$ distributed. The bandwidth is chosen as $h_n = n^{-1/2}$ and the trimming as $e_n = n^{-1/5}$.

**Empirical level** To study the empirical levels of $T_1$ and $T_2$ we choose

$$g(y) = \tilde{g}(y) = \exp\{(y + 3)/(y + 5)\},$$

and we fix a theoretical level $\alpha = 5\%$. Table 1 shows empirical levels of the test under $H_0$. It can be seen that both statistics $T_1$ and $T_2$ provide levels close to the asymptotic value.

### 3.2 Study of the empirical powers

We consider the model where $Y_t$ and $\tilde{Y}_t$ are independent and $\mathcal{N}(0,1)$ distributed. To study the empirical powers of $T_1$ and $T_2$ we consider $g(y) = \exp((y+3)/(y+5))$ and the four following transformations:

$$\tilde{g}_1(y) = \exp((y+3)/(y+5)) + 1, \tilde{g}_2(y) = 2\exp((y+3)/(y+5)),$$
$$\tilde{g}_3(y) = -(y+11)/(y+5), \tilde{g}_4(y) = 4y+5,$$

and we also study local alternatives by considering:

$$\tilde{g}_5(y) = g(y) + \frac{2(y+5)}{n^\beta}.$$

Tables 2-3 present empirical powers for $T_1$ and $T_2$ under fixed and local alternatives, respectively, for a theoretical level $\alpha$ equal to 5%. From Table 2 it appears that the knowledge of the probability densities of $Y$ and $\tilde{Y}$ allows to have more stable statistics that detect more easily the departure from the null hypothesis. Then the test statistic $T_1$ provides better power, particularly for smallest sample size. The test statistic $T_2$ has a low empirical power for $n = 50$; but when the sample size $n$ increases, the empirical power of $T_2$ is similar to that of $T_1$. Table 3 indicates that $T_1$ and $T_2$ provide good power for $\beta \leq 1/2$. For $\beta > 1/2$ the power converges to the theoretical level $\alpha$; this is in accordance with the theoretical result stated in Proposition 2.5.

### 3.3 Real example: Framingham data

We consider the Framingham Study on coronary heart disease described by Carroll et al. [4]. The data consist of measurements of systolic blood pressure (SBP) obtained at two different examinations in 1,615 males on an 8-year follow-up. At each examination, the SBP was measured twice for each individual. The four variables of interest are:
$Y$ = the first SBP at examination 1,
$\tilde{Y}$ = the second SBP at examination 1,
$X$ = the first SBP at examination 2,
$\tilde{X}$ = the second SBP at examination 2.

Our purpose is to examine whether the distribution of the SBP changed during time, and which type of transformation it underwent. Following our notations, we will study the transformation between the distributions of $Y$ and $X$ and also the one between the distributions of $\tilde{Y}$ and $\tilde{X}$. Then we assume that $X = g(Y)$ and $\tilde{X} = \tilde{g}(\tilde{Y})$.

Table 4 indicates that all the distributions of $X$, $Y$, $\tilde{X}$ and $\tilde{Y}$ are skewed to the right and are leptokurtic, $KS$ is the Kolomogorov-Smirnov statistic, the associated p-values are lesser than $2.210^{-6}$ and hence the normality assumption is strongly rejected. Figure 1 represents nonparametric estimations of the probability densities of $X, Y, \tilde{X}$, and $\tilde{Y}$.

From Figure 1 it seems that the distributions of the variables $Y$ and $X$ have a similar shape. However, from Table 4 we observe a noticeable decrease in the mean and an increase in the variance. Based on the nonparametric estimators given in Figure 2 we can postulate

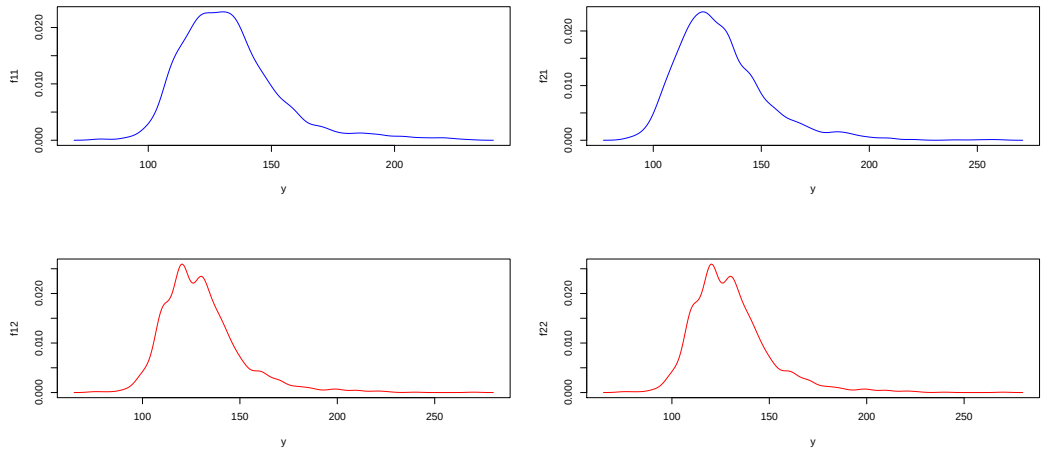


Figure 1: Kernel estimates of the probability densities of $X, Y, \tilde{X}, \tilde{Y}$. In the top panel : f11 (resp. f21) is the Kernel estimate of the density of $Y$ (resp. of $X$). In the bottom panel : f12 (resp. f22) is the Kernel estimate of the density of $\tilde{Y}$ (resp. of $\tilde{X}$).

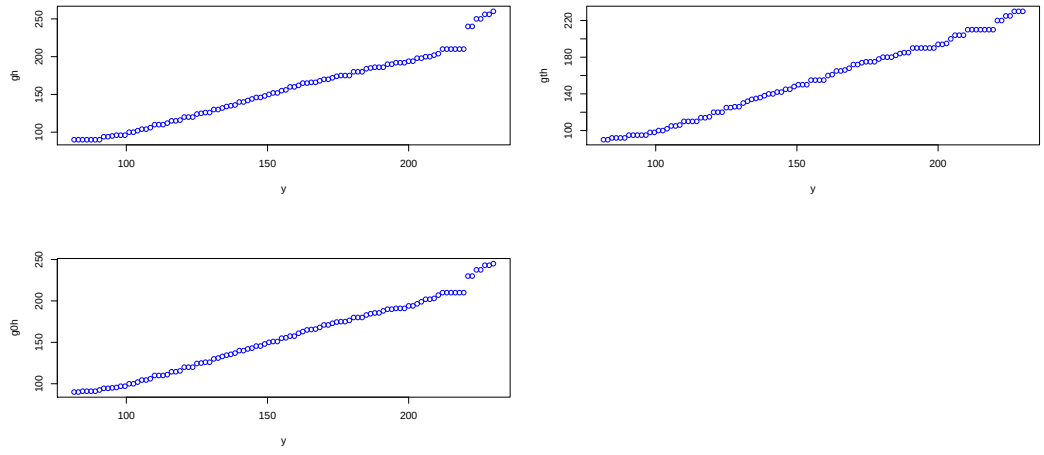


Figure 2: Nonparametric estimators of $g$ and $\tilde{g}$ and the aggregated estimator on the interval $[c, d]$: $gh$ (resp. $gth$ and $g0h$) denotes $\widehat{g}$ (resp. $\widehat{\tilde{g}}$ and $\widehat{g_0}$).

that only the location and the scale are affected by time, therefore, the transformation $g$ is linear; that is, $g(y) = ay + b$. Similarly the distributions of the variables $\tilde{Y}$ and $\tilde{X}$ can be linked by $\tilde{g}(y) = \tilde{a}y + \tilde{b}$. The functions $g$, $\tilde{g}$ are estimated on the interval $[c, d]$ where $c = \max(\min(Y_i), \min(\tilde{Y}_j))$ and $d = \min(\max(Y_i), \max(\tilde{Y}_j))$. These functions are estimated on the grid $y_i = c + (d - c)i/M$, for a given $M$.

By applying our test we obtain a p-value very close to 1, and hence we can consider that $g = \tilde{g}$.

In Figure 2 we observe that all the estimators $\widehat{g}$, $\widehat{\tilde{g}}$ and $\widehat{g_0}$ are approximately linear on the interval $[c, d]$, however in the border (near $c$ and $d$) the approximation is not good. One can observe that they are constants on regions where there are not enough observations. Therefore, to compute the linear approximation of these estimators we consider only the $y_i$ belonging to the interval $[100, 200]$.

The ordinary least squares based on $(y_i, \widehat{g}(y_i))$, $(y_i, \widehat{\tilde{g}}(y_i))$ and $(y_i, \widehat{g_0}(y_i))$, $y_i \in [100, 200]$, $1 \leq i \leq 50$ yields

$$\widehat{g}(y) = 0.9877y + 0.7035, \widehat{\tilde{g}}(y) = 0.9857y + 0.8335 \text{ and } \widehat{g_0}(y) = 0.986y + 0.7685$$

By using a parametric approach, i.e. $\widehat{g_p}(y) = ay + b$, where $a = \text{cov}(X, Y)/\text{var}(Y), b = \overline{X} - a\overline{Y}$, we obtain the following estimators

$$\widehat{g_p}(y) = 0.760y + 33.075, \widehat{\tilde{g}}_p(y) = 0.726y + 36.730,$$

and the common aggregate parametric estimator is given by

$$\widehat{h_{p,0}}(y) = 0.744y + 34.787.$$

To compare the parametric and the nonparametric approaches, we consider the aggregate estimators and we compare the predicted values for the two first moments of $X$ and $\tilde{X}$ with those observed. The predictions of $X$ ( resp. of $\tilde{X}$) are computed by using the observed moments of $Y$ (resp. of $\tilde{Y}$) and the common transformation. Using the parametric approach we get

$$\widehat{m}_X = 0.744 m_Y + 34.787 = 133.590$$
$$\widehat{Var}_X = (0.744)^2 Var(Y) = 232.145.$$

The nonparametric approach yields

$$\widehat{m}_X = 0.9867 m_Y + 0.7685 = 131.8$$
$$\widehat{Var}_X = (0.9867)^2 Var(Y) = 408.04$$

Note that the observed two first moments of $X$ are given by 131.2 and 439.11.
Similarly for the pair $(\tilde{X}, \tilde{Y})$, the parametric predictions are given by

$$\widehat{m}_{\tilde{X}} = 0.744 m_{\tilde{Y}} + 34.787 = 131.656$$
$$\widehat{Var}_{\tilde{X}} = (0.744)^2 Var(\tilde{X}) = 226.933.$$

The nonparametric approach yields

$$\widehat{m}_{\tilde{X}} = 0.9867 m_{\tilde{Y}} + 0.7685 = 129.237,$$
$$\widehat{Var}_{\tilde{X}} = (0.9867)^2 Var(\tilde{Y}) = 399.137$$

Recall that the observed two first moments of $\tilde{X}$ are given by 128.8 and 410.21.
The predictions of the nonparametric model are more close to the observed values, consequently the nonparametric approach seems to be more suitable.

## References


[1] N. Bissantz, L. Dümbgen, H. Holzmann, H and A. Munk, *Non-parametric confidence bands in deconvolution density estimation* J. Roy. Stat. Soc. B, 69 (2007), pp. 483–506.

[2] C. Butucea and C. Matias, *Minimax estimation of the noise level and of the deconvolution density in a semiparametric deconvolution model*, Bernoulli 11 (2005), pp. 309–340.

[3] R.J. Carroll and P. Hall, *Optimal rates of convergence for deconvolving a density*, J. Amer. Statist. Assoc. 83 (1988), pp. 1184–1186.

[4] R.J. Carroll, D. Ruppert, L.A. Stefanski and C. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective*, Second Edition. Chapman Hall, New York, 2006.

[5] L. Cavalier and N. Hengartner *Adaptive estimation for inverse problems with noisy operators*, Inverse Problems 21 (2005), pp. 1345–1361.

[6] F. Comte and T. Rebafka. Adaptive density estimation in the pile-up model involving measurement errors. Available at http://arxiv.org/abs/1011.0592, 2010.

[7] L. Devroye, *Consistent deconvolution in density estimation*, Canad. J. Statist. 17 (1989) pp. 235–239.

[8] P. Diggle and P. Hall, *A Fourier approach to nonparametric deconvolution of a density estimate*, J. Roy. Statist. Soc. Ser.B 55 (1993) pp. 523–531.

[9] S. Efromovich and V. Koltchinskii, *On inverse problems with unknown operators*, IEEE Trans. Inform. Theory 47 (2001), pp. 2876–2893.

[10] J. Fan, *On the optimal rate of convergence for nonparametric deconvolution problems*, Ann. Statist. 19 (1991), pp. 1257–1272.

[11] W. Härdle, Applied Nonparametric Regression. Cambridge Books, Cambridge University Press, 1992.

[12] H. Holzmann, N. Bissantz and A. Munk, *Density testing in a contaminated sample*, J. Multiv. Anal. 98 (2007), pp. 57–75

[13] J. Johannes, S. Van Bellegem and A. Vanhems, *Convergence rates for ill-posed inverse problems with an unknown operator*, Tech. Rep., IDEI Working Paper, 2009.

[14] A. Meister, *Density estimation with normal measurement error with unknown variance*, Statist. Sinica 16 (2006), pp. 195–211.

[15] M.H. Neumann, *Deconvolution from panel data with unknown error distribution*, J. Multivariate Anal. 98 (10) (2007), pp. 1955–1968.

[16] P.K. Sen, *Limiting behavior of regular functionals of empirical distributions for stationary mixing processes* Probability Theory and Related Fields, 25 (1972), pp. 71–82.

[17] X.F. Wang and B. Wang, *Deconvolution Estimation in Measurement Error Models*: The R Package decon. Journal of Statistical Software, 39, 10 (http://www.jstatsoft.org/).

Table 1: Empirical levels of $T_1$ and $T_2$ (in %) for a theoretical level $\alpha = 5\%$ .

| | $n = 50$ | $n = 100$ | $n = 500$ |
|---|---|---|---|
| $T_1$ | 3.9 | 4.75 | 5.49 |
| $T_2$ | 4.68 | 5.52 | 5.42 |

Table 2: Empirical powers of $T_1$ and $T_2$ (in %) for a theoretical level $\alpha = 5\%$.

| | $T_1$ | $T_2$ | $T_1$ | $T_2$ |
|---|---|---|---|---|
| | $\tilde{g}_1$ | $\tilde{g}_1$ | $\tilde{g}_2$ | $\tilde{g}_2$ |
| $n = 50$ | 99.98 | 99.58 | 99.81 | 98.17 |
| $n = 100$ | 99.99 | 99.66 | 99.91 | 98.17 |
| $n = 500$ | 100 | 99.69 | 99.96 | 98.47 |
| | $T_1$ | $T_2$ | $T_1$ | $T_2$ |
| | $\tilde{g}_3$ | $\tilde{g}_3$ | $\tilde{g}_4$ | $\tilde{g}_4$ |
| $n = 50$ | 100 | 100 | 78.59 | 71.47 |
| $n = 100$ | 100 | 100 | 84.31 | 78.41 |
| $n = 500$ | 100 | 100 | 92.42 | 92.07 |

Table 3: Empirical powers of $T_1$ and $T_2$ (in %) for a theoretical level $\alpha = 5\%$ under local alternative $\tilde{g}_5$.

| | $T_1$ | $T_2$ | $T_1$ | $T_2$ | $T_1$ | $T_2$ |
|---|---|---|---|---|---|---|
| | $\beta = 1/4$ | $\beta = 1/4$ | $\beta = 1/2$ | $\beta = 1/2$ | $\beta = 4$ | $\beta = 4$ |
| $n = 50$ | 99.85 | 97.06 | 99.64 | 96.90 | 4.19 | 4.77 |
| $n = 100$ | 99.85 | 97.30 | 99.71 | 97.02 | 4.77 | 5.72 |
| $n = 500$ | 99.94 | 97.85 | 99.82 | 97.29 | 5.36 | 5.28 |

Table 4: Descriptive statistics of Framingham data

| $Y$ | $X$ |
|---|---|
| Min. 1st Qu. Median Mean 3rd Qu. Max. | Min. 1st Qu. Median Mean 3rd Qu. Max. |
| 80.0 120.0 130.0 132.8 142.0 230.0 | 88.0 118.0 128.0 131.2 142.0 260.0 |
| Var. Skewness. Kurtosis. KS. | Var. Skewness. Kurtosis. KS. |
| 419.12 1.27 7.79 0.0119 | 439.11 1.39 6.65 0.1125 |
| $\tilde{Y}$ | $\tilde{X}$ |
| Min. 1st Qu. Median Mean 3rd Qu. Max. | Min. 1st Qu. Median Mean 3rd Qu. Max. |
| 75.0 118.0 128.0 130.2 140.0 270.0 | 85.0 115.0 125.0 128.8 138.0 270.0 |
| Var. Skewness. Kurtosis. KS. | Var. Skewness. Kurtosis. KS. |
| 409.97 1.46 7.25 0.1171 | 410.21 1.47 7.10 0.1117 |